Hypothesis Tests for Evaluating Numerical Precipitation Forecasts

THOMAS M. HAMILL

Advanced Studies Program, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 18 May 1998, in final form 9 November 1998)

ABSTRACT

When evaluating differences between competing precipitation forecasts, formal hypothesis testing is rarely performed. This may be due to the difficulty in applying common tests given the spatial correlation of and nonnormality of errors. Possible ways around these difficulties are explored here. Two datasets of precipitation forecasts are evaluated, a set of two competing gridded precipitation forecasts from operational weather prediction models and sets of competing probabilistic quantitative precipitation forecasts from model output statistics and from an ensemble of forecasts. For each test, data from each competing forecast are collected into one sample for each case day to avoid problems with spatial correlation. Next, several possible hypothesis test methods are evaluated: the paired *t* test, the nonparametric Wilcoxon signed-rank test, and two resampling tests. The more involved resampling test methodology is the most appropriate when testing threat scores from nonprobabilistic forecasts. The simpler paired *t* test or Wilcoxon test is appropriate to use in testing the skill of probabilistic forecasts evaluated with the ranked probability score.

1. Introduction

Improving the accuracy of quantitative precipitation forecasts is a primary goal of the National Centers for Environmental Prediction and the meteorological research community (Fritsch et al. 1998). Frequently, whether or not to update an operational model's physics is based at least in part on whether precipitation forecast skill appears to be improved (e.g., Rogers et al. 1995; Rogers et al. 1996). This may involve the subjective evaluation of the models during significant precipitation events and/or comparison of statistical measures of precipitation forecast quality.

A difference in ''threat score'' is often provided as evidence of difference in forecast skill between competing gridded forecasts. These scores are generated from a contingency table of hits, misses, false alarms, and correct no forecasts. Currently, the most commonly cited scores are the equitable threat score, or ETS [Schaefer (1990); also known as the Gilbert skill score] and the bias, or BIA. The ETS has the desirable property that a perfect forecast has skill 1.0, and constant or random forecasts have skill 0.0. A bias of 1.0 indicates that events are forecast with the same frequency as they occur. For probabilistic forecasts, Brier scores (Brier 1950; Wilks 1995), or ranked probability skill scores (RPSS; Epstein 1969; Murphy 1971; Daan 1985; Wilks 1995) are commonly used to evaluate forecasts.

When evaluating model differences, the two competing forecast systems should always be compared over a wide range of weather conditions. If one has an improved ETS or RPSS, this may provide some evidence that this forecast system is superior. However, assessing the confidence in such an assessment may be difficult for a host of reasons. First, comparisons of common threat scores like the ETS are suspect unless the biases of competing forecasts are similar; typically, the forecast with the larger bias (the wetter forecast) tends to have a higher ETS than if the two models had the same bias (Mason 1989). Second, a user wishing to conduct a hypothesis test of the statistical significance of skill score differences may wonder if established hypothesis tests may be used. Precipitation forecast errors are often non-normally distributed and have spatially and/or temporally correlated error (Wilks 1997; Livezey and Chen 1983); the effective number of independent samples is much less than the total number of grid points. Hence, the tester will be misled if his test methodology treats individual grid points as independent samples.

Hypothesis testing is typically performed using single numbers as samples, such as daily threat scores. However, a final overall threat score is normally generated from a sum of daily contingency tables, not from an average of daily threat scores. If the daily contingency tables are reduced to daily threat scores for the purpose of hypothesis testing, the data is in a form appropriate for the use of common hypothesis tests such as the

Corresponding author address: Dr. Thomas M. Hamill, NCAR/MMM, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: hamill@ucar.edu

paired *t* test, yet the use of such tests may be inappropriate. As discussed more in section 3a, a daily threat score sample may change dramatically with only small changes in the partitioning of the counts in the contingency tables. Unfortunately, common hypothesis test results may be unduly sensitive to these changes. At high precipitation many case days also may have no rainfall forecast and/or no rainfall observed above this threshold in both competing models. In this event, the sample population of daily threat scores may be dominated by zeros for both forecasts. This may violate the assumption of normality of errors that are assumed in hypothesis tests such as the paired *t* test. Whether such a test is still appropriate to use is not well understood.

For all these reasons, formal hypothesis testing of competing forecast models has largely been neglected, and decisions on whether to make model changes are often made in ignorance of whether or not the differences are statistically significant. Guidance is needed on when simple hypothesis tests are appropriate to use and when and if more complex or computationally expensive tests are required.

Modem, computer-based methods of hypothesis testing have been developed over the last few decades to permit hypothesis testing in situations where it is not clear whether classic tests will work properly. Methods based on the resampling technique (Diaconis and Efron 1983; Livezey and Chen 1983; Good 1994; Wilks 1995; Briggs and Levine 1998) will be demonstrated in this paper and compared against two more established techniques. The underlying principle or resampling is simple: use the computer to build a distribution consistent with one's null hypothesis by repeated random sampling from the collected data, and assess the significance of the test from the ranking of the observed test statistic in this distribution.

The paper will be organized as follows. Section 2a describes the nonprobabilistic and probabilistic precipitation test data to be used to demonstrate candidate hypothesis test techniques. Sections 2b and 2c provide a brief review of threat scores for evaluating nonprobabilistic precipitation forecasts and the RPSS for evaluating probabilistic quantitative precipitation forecasts (PQPFs). Section 3a explores the problems and solutions unique to applying hypothesis tests to precipitation datasets. Sections 3b and 3c describe the candidate hypothesis tests for nonprobabilistic and probabilistic forecasts, respectively. Section 4 demonstrates the application of these test methods to sample datasets and discusses the relative merits of each. Section 5 provides conclusions and recommendations.

Since precipitation forecasts in the United States are still commonly issued in the English units of inches, this convention will be used (1.0 in. = 25.4 mm).

2. Verification measures and data

a. Forecast and verification data

Hypothesis test methodologies will be demonstrated here for both nonprobabilistic and probabilistic forecasts. To demonstrate hypothesis tests for nonprobabilistic forecasts, approximately 5 months of gridded forecasts and verification data from the Nested Grid Model (NGM; Hoke et al. 1989; Petersen et al. 1991) and the Meso Eta 29-km model (Black 1994) will be compared. A total of 160 days of data were available from October 1997 to March 1998. The 24-h total precipitation observations were taken from the River Forecast Center database; each observation was assigned to its nearest grid box, and the observed gridpoint precipitation represents the average of all observations assigned to that box. Grid points where no observations were available were excluded. Gridded forecasts of rainfall were remapped and integrated to a common 80-km grid in a method that conserves total water (Mesinger 1996; M. Baldwin 1998, personal communication). Generally, grid points with valid observations were limited to the conterminous United States, with more valid grid points in the east than in the west. The hypothesis test methodologies will be demonstrated at the 0.01-, 0.10-, 0.25-, 0.50-, 0.75-, 1.00-, 1.50-, and 2.00-in. thresholds.

For probabilistic precipitation forecasts, model output statistics (MOS; Carter et al. 1989) PQPFs, and prototype Eta/Regional Spectral Model (RSM) ensemble categorical PQPFs will be evaluated using the same 13 case days and verification data from 1995–96 described in Hamill and Colucci (1998). On each case day, the verification data (12-hourly station precipitation totals) and station forecasts were jointly available at approximately 300 sites. Here, we will focus on the comparisons of categorical PQPFs generated 1) by MOS and 2) by fitting a gamma distribution to the ensemble mean and subsequent categorization of probabilities, as described in Hamill and Colucci (1998). This latter forecast method will be denoted $\Gamma(\overline{\text{Eta}})$. Forecast probabilities and observations were assigned to the MOS precipitation categories. These categories are $0 \le V < 0.01$ in., $0.01 \le V < 0.10$, $0.10 \le V < 0.25$, $0.25 \le V < 0.5$, $0.5 \le V < 1.0$, $1.0 \le V < 2.0$, and $V \ge 2.0$, where V is the verification amount in inches.

b. Equitable threat score and bias

These scores are useful for evaluating nonprobabilistic, gridded precipitation forecasts. They are generated as follows. For a given precipitation threshold, forecasts are partitioned into a contingency table of four mutually exclusive and collectively exhaustive events: (*a*) the number of locations with both forecast and verification greater or equal than the threshold, that is, ''hits;'' (*b*) number of locations with forecast at or above the threshold and verification below, or ''false alarms;'' (*c*) number of locations with forecast below and verification at or above the threshold, or ''misses;'' and (*d*) forecast and verification both below the threshold. This is illustrated in Table 1. The ETS is defined by

TABLE 1. Contingency table of possible events.

		Observed	
		Yes	No
Forecast	Yes	*a*	*b*
	No	*c*	*d*

$$\text{ETS} = \frac{a - a_r}{a + b + c - a_r}, \tag{1}$$

where a_r is the expected number of correct forecasts above the threshold in a random forecast, where forecast yeses/nos are independent of observation yeses/nos, defined by

$$a_r = \frac{(a + b)(a + c)}{a + b + c + d}. \tag{2}$$

The bias is the ratio of the number of yes forecasts issued divided by the number of yes observed:

$$\text{BIA} = \frac{a + b}{a + c}. \tag{3}$$

c. Ranked probability skill score

The ranked probability score (RPS; Epstein 1969; Murphy 1971) is a single-number statistic that indicates the quality of a set of probabilistic forecasts for a set of ordered categories. Using a forecast distribution vector of precipitation probabilities **y** with k categories, a cumulative distribution vector **Y** is defined with components

$$\mathbf{Y}_m = \sum_{j=1}^{m} y_j, \qquad m = 1, \ldots, k. \tag{4}$$

Similarly, from the vector of the observations **o**, a cumulative distribution vector **O** is also generated:

$$\mathbf{O}_m = \sum_{j=1}^{m} o_j, \qquad m = 1, \ldots, k, \tag{5}$$

where $o_j = 1$ if precipitation occurred in the jth category and is zero otherwise. The RPS is the squared difference between the forecast's cumulative distribution vector and the observed cumulative distribution vector,

$$\text{RPS} = \sum_{m=1}^{k} (\mathbf{Y}_m - \mathbf{O}_m)^2. \tag{6}$$

The RPSS measures the fractional improvement of the average of rank probability score over all forecasts relative to the average score of a reference forecast, $\overline{\text{RPS}}_R$ (Wilks 1995):

$$\text{RPSS} = 1 - \frac{\overline{\text{RPS}}}{\overline{\text{RPS}}_R}. \tag{7}$$

The overbar denotes an average over all forecasts. The reference forecast is commonly derived from climatology, a persistence forecast, or an established forecast system.

3. Hypothesis test methodologies

a. Test design principles

We first outline some of the issues that must be addressed in designing an appropriate hypothesis test for precipitation forecast data.

1) SPATIAL CORRELATION OF ERROR

Because of spatial correlation of forecast error, individual gridbox elements cannot be considered independent samples. If the forecast missed a rain event for Washington, D.C., it also was likely that nearby Baltimore, Maryland, missed the rain event as well. Hence, a single daily sample statistic is calculated from a sum or average over all model grid points where valid observations were taken. Subdividing a daily gridded forecast into even slightly smaller units resulted in correlated error among the subdivisions, violating the assumption of independence of samples. To illustrate this, for a set of gridded model forecasts and observations over the conterminous United States for a 160-day period (section 2a), the verification region was subdivided into four smaller blocks with approximately equal numbers of grid points with observations. The Spearman rank correlation of ETS was often significantly greater than zero among the subblocks, as indicated in Table 2.

2) PAIRING OF SAMPLES

Because the performance of both competing forecast schemes on a given day are related to the synoptic conditions on that day, the hypothesis test should treat the forecast data as paired. On days where the weather is dominated by high pressure, both forecast models are likely to correctly forecast large areas of no precipitation, but on stormy days both forecasts are likely to exhibit generally higher than normal error. Pairing of the data reduces the variance and results in a more powerful hypothesis test.

3) SERIAL CORRELATION

Whether serial (temporal) correlation of error need be addressed was not immediately obvious. If today's forecast was too rainy, can we expect tomorrow's forecast to be too rainy as well? To examine this, ETS and BIA scores were calculated each day and each precipitation threshold from a set of NGM and Meso Eta 48-h forecasts during December 1997, a sequence of 31 continuous days where forecast and verification data was available. A lag-one Spearman rank correlation analysis was performed to see if there was a statistically signif-

TABLE 2. Spearman rank correlation of daily ETS among subblocks of model forecast domain using Meso Eta data from a 160-day period in 1997–98. Subdomains are northeast (NE), northwest (NW), southeast (SE), and southwest (SW) United States. Two-sided *p* values of the significance of the deviation from zero are provided as well.

Threshold (in.)	Block 1	Block 2	Spearman rank corr.	*p* value
0.01	SW	NW	0.26	0.001
	SW	SE	0.23	0.003
	SW	NE	−0.07	0.360
	NW	SE	−0.02	0.783
	NW	NE	0.02	0.770
	SE	NE	0.29	<0.001
0.10	SW	NW	0.29	<0.001
	SW	SE	0.24	0.002
	SW	NE	−0.01	0.869
	NW	SE	−0.03	0.670
	NW	NE	−0.02	0.765
	SE	NE	0.25	0.001
0.25	SW	NW	0.24	0.001
	SW	SE	0.29	<0.001
	SW	NE	0.80	<0.001
	NW	SE	0.81	<0.001
	NW	NE	0.04	0.591
	SE	NE	0.22	0.004
0.50	SW	NW	0.32	<0.001
	SW	SE	0.36	<0.001
	SW	NE	0.45	<0.001
	NW	SE	0.18	0.022
	NW	NE	0.31	<0.001
	SE	NE	0.58	<0.001
1.00	SW	NW	0.63	<0.001
	SW	SE	0.42	<0.001
	SW	NE	0.99	<0.001
	NW	SE	0.33	<0.001
	NW	NE	0.94	<0.001
	SE	NE	0.89	<0.001

icant correlation. A mixture of both positive and negative correlations was observed. Table 3 reports the observed two-sided significance of the deviation of the rank correlation from zero. As shown in the table, there is no evidence to indicate precipitation forecasts a day apart exhibit serial correlation of threat score. Hence, we assume each case day may effectively be treated as independent from prior and subsequent days. Whether this holds for forecasts 12 h apart is not known.

4) DIFFERING MODEL BIASES

As noted earlier, comparisons of the ETS from competing forecasts may be misleading if their biases are dissimilar (Mason 1989). For example, given two forecast systems with similar ETS and BIA, the ETS of one forecast system may be inflated by adding a constant precipitation amount if doing so tends to preferentially increase the number of hits. This deficiency is illustrated in Figs. 1a–c. Figure 1a shows a hypothetical forecast and observed precipitation pattern. Normally, to calculate the ETS at a threshold, say 0.75 in., the forecast

TABLE 3. Two-sided significance (p value) of the deviation from zero of lag-1 rank correlations of ETS and BIA from 48-h Meso Eta and NGM forecasts using the data described in section 2c.

Threshold (in.)	Meso Eta 48-h ETS	Meso Eta 48-h BIA	NGM 48-h ETS	NGM 48-h BIA
0.01	0.86	0.69	0.63	0.40
0.10	0.85	0.68	0.58	0.91
0.25	0.65	0.61	0.42	0.33
0.50	0.65	0.32	0.62	0.14
0.75	0.43	0.88	0.73	0.15
1.00	0.75	0.17	0.36	0.74
1.50	0.24	0.34	0.69	0.77
2.00	0.56	0.29	0.56	0.56

areas above and below the threshold are calculated using the 0.75-in. forecast contour. However, if 0.10 in. is added to the forecast at every grid point, then the original 0.65-in. forecast contour may now be considered the forecast threshold. Proceeding in this manner, Figs. 1b,c illustrate the ETS and BIA over the range of choices of forecast threshold while holding the verification threshold fixed at 0.75 in. As shown, in this instance the ETS is maximized if the 0.35-in. contour is chosen as the threshold. That is, a forecast with 0.4 in. arbitrarily added to the forecast at all grid points would score much higher in ETS than forecasts with none or less of this ''gaming'' of the precipitation field. Hence, when the ETS of competing forecasts are evaluated, the forecast with the higher BIA is likely to have a higher ETS than if the forecast were adjusted to have the same BIA as its competitor. This also complicates application of the hypothesis test, for a higher ETS by one model may not be unambiguously attributed to higher skill if model biases differ. Even though forecast models are certainly not designed with the gaming of ETS in mind, if two competing models have different biases, this effect should be considered. Hypothesis tests will be demonstrated in section 4 with and without a correction for this effect.

5) SENSITIVITY TO SMALL CHANGES IN CONTINGENCY TABLE POPULATION

For rare events such as heavy precipitation, only one element of the forecast table *d* in Table 1, the ''correct no'' element, may be highly populated. Because the other elements are sparsely populated, a small change in the population of the elements can cause a large change in daily threat scores. To illustrate this, consider a scenario of two subsequent days, one dry and the next one moist. Perhaps a contingency table (*a, b, c, d*) over 1000 grid points on the first day was populated with (0, 2, 1, 997) for the NGM forecast and (0, 1, 2, 997) for the Meso Eta. The respective biases for the two models are $(0 + 2)/(0 + 1) = 2$ and $(0 + 1)/(0 + 2) = 0.5$. On the next day, the contingency table elements are much more evenly populated, say (50, 80, 80, 790) and (60, 70, 90, 880), yielding biases of $(50 + 80)/(50 +$

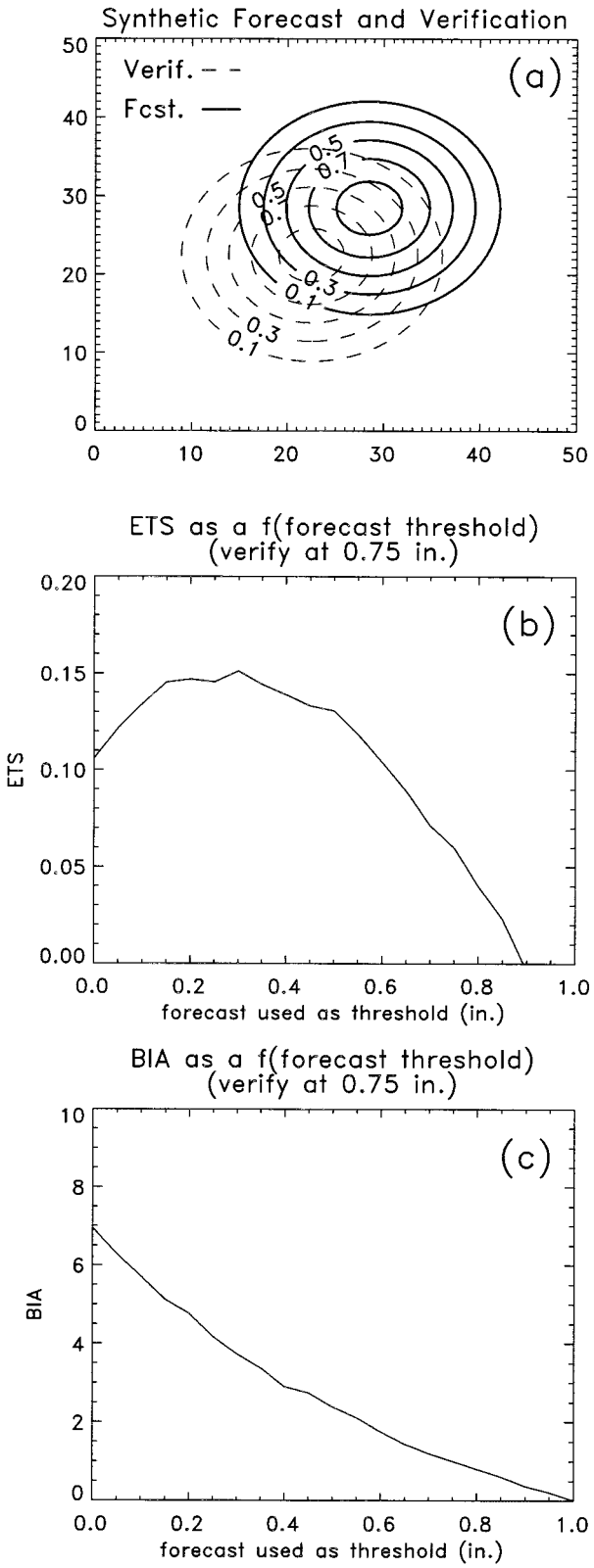

FIG. 1. Illustration of how the equitable threat score may be inflated for forecasts with higher bias. (a) Synthetic verification (dashed) and forecast (solid) of precipitation. (b) ETS as a function of the forecast used for the threshold in the case where the verification threshold is 0.75 in. (c) Bias as a function of the forecast threshold.

80) = 1.0 and (60 + 70)/(60 + 90) = 0.867, respectively. The bias difference between models on the first day is 2.0 − 0.5 = 1.5, but this sample difference is very sensitive to a redistribution of a single count among the contingency table elements. Conversely, changing the population of contingency tables in the same way on the second day will not have nearly as much effect. Ideally, the hypothesis test should be designed to not be sensitive to these small changes.

b. Test methodologies for precipitation threat scores

We first outline the most complex methodology, a technique based on resampling. Accessible introductions to hypothesis testing via resampling are provided by Diaconis and Efron (1983) and Wilks (1995). More detailed information on its proper use can be found in Hall and Wilson (1991), Efron and Tabshirani (1993), and Good (1994). The methodology properly addresses the test design problems discussed in section 3a.

The null hypotheses for the resampling tests are that the differences in ETS and BIA between the two competing forecasts M1 and M2 are zero, computed from a sum of daily contingency table samples over all case days; that is,

$$H_0: \quad \mathrm{ETS}_{\mathrm{M1}} - \mathrm{ETS}_{\mathrm{M2}} = 0.0, \qquad \mathrm{BIA}_{\mathrm{M1}} - \mathrm{BIA}_{\mathrm{M2}} = 0.0, \tag{8}$$

and the alternative hypotheses

$$H_A: \quad \mathrm{ETS}_{\mathrm{M1}} - \mathrm{ETS}_{\mathrm{M2}} \neq 0.0, \qquad \mathrm{BIA}_{\mathrm{M1}} - \mathrm{BIA}_{\mathrm{M2}} \neq 0.0. \tag{9}$$

Assume a two-sided test with significant level $\alpha = 0.05$. Next we form a test statistic and a resampled distribution consistent with the null hypothesis. Each daily sample from each model is a vector of the contingency table elements:

$$\mathbf{x}_{i,j} = (a, b, c, d)_{i,j}, \qquad i = 1, 2, \quad \text{and} \quad j = 1, \ldots, n, \tag{10}$$

where n is the number of case days. Here i is the indicator of the forecast model, and j is the number of the case day. The test statistic

$$\left(\widehat{\mathrm{ETS}}_{\mathrm{M1}} - \widehat{\mathrm{ETS}}_{\mathrm{M2}}\right) \quad \text{or} \quad \left(\widehat{\mathrm{BIA}}_{\mathrm{M1}} - \widehat{\mathrm{BIA}}_{\mathrm{M2}}\right)$$

is calculated using (1)–(3) after summing contingency table elements for each model over all case days:

$$\widehat{(a, b, c, d)}_{\mathrm{M1}} = \sum_{k=1}^{n} \mathbf{x}_{1,k}, \tag{11}$$

and

$$\widehat{(a,\ b,\ c,\ d)}_{\mathrm{M2}} = \sum_{k=1}^{n} \mathbf{x}_{2,k}. \tag{12}$$

Resampled test statistics consistent with the null hypothesis are generated after randomly choosing either one or the other model on each day and summing contingency table elements. Let I_j be an random indicator variable equally likely to take on the value 1 or 2, with $j = 1, \ldots, n$. To calculate the resampled test statistic, generate the n random samples of I and form a resampled sum of the shuffled vectors of contingency table elements over all case days:

$$\widehat{(a,\ b,\ c,\ d)}_1^* = \sum_{k=1}^{n} \mathbf{x}_{I_k,k}. \tag{13}$$

Form another sum of contingency table elements using the model data not selected for the first sum in (13):

$$\widehat{(a,\ b,\ c,\ d)}_2^* = \sum_{k=1}^{n} \mathbf{x}_{(3-I_k),k}. \tag{14}$$

From $\widehat{(a,\ b,\ c,\ d)}_1^*$ and $\widehat{(a,\ b,\ c,\ d)}_2^*$ the resampled test statistics

$$\left(\widehat{\mathrm{ETS}}_1^* - \widehat{\mathrm{ETS}}_2^*\right) \quad \text{or} \quad \left(\widehat{\mathrm{BIA}}_1^* - \widehat{\mathrm{BIA}}_2^*\right)$$

are then calculated using (1)–(3). This process is repeated many times (here, 1000) to build a null distribution.

The hypothesis of difference in ETS is finally tested by determining the location of $(\widehat{\mathrm{ETS}}_{\mathrm{M1}} - \widehat{\mathrm{ETS}}_{\mathrm{M2}})$ in the resampled distribution of $(\widehat{\mathrm{ETS}}_1^* - \widehat{\mathrm{ETS}}_2^*)$, and similarly for bias. Formally, using the resampled distribution we compute numbers $\widehat{t_L}$ and $\widehat{t_U}$ such that

$$\mathrm{Pr}^*\left[\left(\widehat{\mathrm{ETS}}_1^* - \widehat{\mathrm{ETS}}_2^*\right) < \widehat{t_L}\right] = \frac{\alpha}{2}, \quad \text{and}$$

$$\mathrm{Pr}^*\left[\left(\widehat{\mathrm{ETS}}_1^* - \widehat{\mathrm{ETS}}_2^*\right) < \widehat{t_U}\right] = 1 - \frac{\alpha}{2}, \tag{15}$$

where Pr* represents probabilities calculated from this distribution. Then H_0 is rejected if

$$(\widehat{\mathrm{ETS}}_{\mathrm{M1}} - \widehat{\mathrm{ETS}}_{\mathrm{M2}}) < \widehat{t_L} \quad \text{or} \quad (\widehat{\mathrm{ETS}}_{\mathrm{M1}} - \widehat{\mathrm{ETS}}_{\mathrm{M2}}) > \widehat{t_U}.$$

As a possible alternative to the resampling methodology, the paired t test or Wilcoxon signed-rank test may be used. Unlike the resampling test, the paired t test requires a vector of the daily differences in ETS or BIA and not contingency table elements. Specifically, given n case days, there are n sample differences in the scores between the competing models. The average threat scores $\overline{\mathrm{ETS}}$ and $\overline{\mathrm{BIA}}$ and their sample standard deviations s_{ETS} and s_{BIA} are calculated from these daily differences. The paired t test assumes $\overline{\mathrm{ETS}}/(s_{\mathrm{ETS}}/\sqrt{n})$ and $\overline{\mathrm{BIA}}/(s_{\mathrm{BIA}}/\sqrt{n})$ are distributed as t variables with $n - 1$ degrees of freedom, and the location of the sample statistic is compared against this distribution. As mentioned previously, test results may be unduly sensitive to small changes in the population of contingency table elements on dry days, and the test statistic is different from the method in which the ETS is itself calculated; the average of daily ETS scores is not necessarily equal to the ETS generated from a sum of daily contingency table elements.

The nonparametric Wilcoxon signed-rank test (Wilks 1985) is also considered here; see Sprent (1989) for a more lengthy discussion of the application of this test when paired sample differences are zero or tied. In this test, given the vector $\mathbf{q}$ of n daily differences between model threat scores, we form a new vector, $\mathbf{z}$, the sorted absolute values of the elements of $\mathbf{q}$. A vector of ranks of $\mathbf{z}$ is generated, indicating the ranking of each element from lowest to highest. Denote this array $\mathbf{t}$. If elements in $\mathbf{z}$ are of equal value, each corresponding element in $\mathbf{t}$ is assigned the same average rank. For example, if $\mathbf{q} = [-1, 1, 0, 0, 3, -4]$, then $\mathbf{z} = [0, 0, 1, 1, 3, 4]$ and $\mathbf{t} = [1.5, 1.5, 3.5, 3.5, 5, 6]$. Now let d_0 equal the number of daily differences in $\mathbf{q}$ equal to zero and denote $\mathrm{d_i}$ as the number of ties in nonsigned ranks other than zero, $i = 1, 2, \ldots, r$, where there are r sets of different ties in the vector of sorted ranks. In our example, $d_0 = 2$, $r = 1$, and $d_1 = 2$. The ranks in $\mathbf{t}$ are then ''signed'' based on whether the original difference was greater than, equal to, or less than zero; $a + 1$, 0, and -1 are multiplied by the ranks, respectively. Denote $\mathbf{u}$ as the signed ranks of $\mathbf{t}$; here $\mathbf{u} = [0, 0, -3.5, 3.5, 5, -6]$. The positive signed ranks U^+ are then summed ($=8.5$ here), and under the null hypothesis the distribution of positive ranks is approximately Gaussian, with mean

$$\mu = \frac{n(n + 1) - d_0(d_0 + 1)}{4}. \tag{16}$$

The standard deviation under the null hypothesis is

$$\sigma = \left[\frac{n(n + 1)(2n + 1) - d_0(d_0 + 1)(2d_0 + 1)}{4} - \frac{\sum_{i=1}^{r} (d_i^3 - d_i)}{48}\right]^{1/2}. \tag{17}$$

The location of the sum of the ranks of positive differences is compared to this distribution; that is, the probability of exceeding the Z score $(U^+ - \mu)/\sigma$ is calculated. This method, too, operates on daily threat scores and thus may have the same drawbacks discussed with the paired t test.

To provide a more direct comparison against the paired t test and Wilcoxon test, another resampling test was designed following the same basic principles of permuting the choice of models on each case day formulated in Eqs. (10)–(15). However, for this resampling test, a daily ETS is calculated for each model from the contingency table elements and the resampling method then permutes the daily ETS samples rather than the vector of contingency table elements. The sample sta-

TABLE 4. Illustration of how different daily statistics may be used as input to the hypothesis tests. The paired *t* test may use the difference in Σ RPS between columns 2 and 3, which emphasizes case days 4 and 10, with high precipitation. The test may also be performed on the skill score difference (relative to 0.0 for MOS), which will emphasize case day 8, with low precipitation. The former is preferred since it is consistent with the method by which an RPSS is calculated in Eq. (7). Skill score from total = −0.023. Average of daily skill scores = −0.078.

Case day	Σ $RPS_{\Gamma(Eta)}$	Σ RPS_{MOS}	Sample size	Skill score
1	69.66	65.13	324	−0.069
2	64.09	77.61	331	0.174
3	63.75	62.42	261	−0.021
4	116.75	91.33	299	−0.278
5	43.34	35.82	318	−0.210
6	43.30	36.40	304	−0.189
7	62.56	63.93	295	0.021
8	28.12	18.65	290	−0.507
9	58.17	57.30	299	−0.015
10	92.66	112.77	288	0.178
11	20.60	18.76	297	−0.097
12	103.80	111.16	301	0.065
13	46.37	43.33	293	−0.070
Total	813.21	794.61	3900	

tistic is the average of daily differences in ETS between models, which is compared to a resampled null distribution of daily ETS differences. This test may be expected to exhibit the same drawbacks as the paired *t* test and Wilcoxon test.

c. Methodology for RPSS

The resampling methodology for RPSS is first described. As with the threat scores, we again make the conservative assumption that the RPSSs may be correlated among verification locations on the same day, so all locations on a given day are treated as a grouped entity. Hypothesis testing here will be performed on the RPS summed over all forecast locations rather than using the daily average RPSS. As with threat scores, this is done so the sample statistic is consistent with the method in which the overall RPSS is calculated (Table 4). The null hypothesis shows no difference in average RPS between competing models, generated from a sum over all observation locations on all days:

$$H_0: \qquad \overline{\mathrm{RPS}}_{\mathrm{M1}} - \overline{\mathrm{RPS}}_{\mathrm{M2}} = 0 \tag{18}$$

and the alternative hypothesis

$$H_A: \qquad \overline{\mathrm{RPS}}_{\mathrm{M1}} - \overline{\mathrm{RPS}}_{\mathrm{M2}} \neq 0. \tag{19}$$

Let us assume that an RPS has been calculated for each model and at each observation location on each case day. The resampling test is most computationally efficient if the RPS scores at individual locations within each case day are summed and resampling permutes the daily sums. Assume there are n total verification locations over the m case days and n_q verification locations on the qth case day. Hence

$$n = \sum_{q=1}^{m} n_q. \tag{20}$$

Further, assume the observation locations are numbered, so the first location on the second case day is numbered $n_1 + 1$, the first on the third day is numbered $n_1 + n_2 + 1$, and so on. Define $n_0 = 0$. Let individual RPS scores be indexed by their model i, $i = 1, 2$, and their observation number k, where $k = 1, \ldots, n$. Hence, $\mathrm{RPS}_{i,k}$ denotes kth score for the ith model. Also, let $R_{i,j}$ be the sum of RPSs of model i on the jth case day, where $i = 1, 2$ and $j = 1, \ldots, m$, where m is the number of case days. The sum of RPS for model i on the jth case day is thus

$$R_{i,j} = \sum_{k=1+\sum_{q=0}^{j-1} n_q}^{\sum_{p=1}^{j} n_p} \mathrm{RPS}_{i,k.} \tag{21}$$

For example, $R_{1,1}$ sums model 1 RPS scores from 1 to n_1, and $R_{1,2}$ from $n_1 + 1$ to $n_1 + n_2$, and so on. The sample average RPSs for models 1 and 2 are then

$$\widehat{\overline{\mathrm{RPS}}}_{\mathrm{M1}} = \frac{\sum_{k=1}^{m} R_{1,k}}{\sum_{k=1}^{m} n_k} \tag{22}$$

and

$$\widehat{\overline{\mathrm{RPS}}}_{\mathrm{M2}} = \frac{\sum_{k=1}^{m} R_{2,k}}{\sum_{k=1}^{m} n_k}, \tag{23}$$

and our test statistic is $(\widehat{\overline{\mathrm{RPS}}}_{\mathrm{M2}} - \widehat{\overline{\mathrm{RPS}}}_{\mathrm{M1}})$.

We now generate a resampled null distribution. As with threat scores, we generate a random indicator variable I_j, $j = 1, \ldots, m$, taking on a value of 1 or 2 with equal probability. The indicator variable is used to randomly select either one or the other model on each case day to develop a resampled sum. Specifically, a resampled statistic

$$\left(\widehat{\overline{\mathrm{RPS}}}_1^* - \widehat{\overline{\mathrm{RPS}}}_2^*\right)$$

is generated using

$$\widehat{\overline{\mathrm{RPS}}}_1^* = \frac{\sum_{k=1}^{m} R_{I_k,k}}{\sum_{k=1}^{m} n_k} \tag{24}$$

and

$$\widehat{\overline{\mathrm{RPS}}}_2^* = \frac{\sum_{k=1}^{m} R_{(3-I_k),k}}{\sum_{k=1}^{m} n_k}. \tag{25}$$

As before, the hypothesis is tested by calculating $\widehat{t_L}$ and $\widehat{t_U}$ such that

$$\Pr^*\left[\left(\widehat{\overline{\mathrm{RPS}}}_2^* - \widehat{\overline{\mathrm{RPS}}}_1^*\right) < \widehat{t_L}\right] = \frac{\alpha}{2} \quad \text{and}$$

$$\Pr^*\left[\left(\widehat{\overline{\mathrm{RPS}}}_2^* - \widehat{\overline{\mathrm{RPS}}}_1^*\right) < \widehat{t_U}\right] = 1 - \frac{\alpha}{2}, \qquad (26)$$

and H_0 is rejected if

$$\left(\widehat{\overline{\mathrm{RPS}}}_{\mathrm{M2}} - \widehat{\overline{\mathrm{RPS}}}_{\mathrm{M1}}\right) < \widehat{t_L} \quad \text{or} \quad \left(\widehat{\overline{\mathrm{RPS}}}_{\mathrm{M2}} - \widehat{\overline{\mathrm{RPS}}}_{\mathrm{M1}}\right) > \widehat{t_U}.$$

For comparison, the paired t test and Wilcoxon signed-rank test were again performed. The input to the paired t test and Wilcoxon test was a vector of differences in total RPS on each case day, that is, $R_{1,j} - R_{2,j}$ in Eq. (21).

4. Results

a. Results for nonprobabilistic forecasts

We demonstrate first the resampling methodology for ETS and BIA from Eqs. (10)–(15) applied to the data from section 2a. Figure 2 shows the comparative ETS and BIA of NGM and Meso Eta forecasts as well as a confidence interval referenced to the Meso Eta forecast. That is, for the ETS, the distance of the error bars from the Meso Eta forecast is the t such that

$$\Pr^*\left[\left(\widehat{\mathrm{ETS}}_1^* - \widehat{\mathrm{ETS}}_2^*\right) > \hat{t}\right] = \frac{\alpha}{2}, \quad \text{and}$$

$$\Pr^*\left[\left(\widehat{\mathrm{ETS}}_1^* - \widehat{\mathrm{ETS}}_2^*\right) > \hat{t}\right] = 1 - \frac{\alpha}{2}.$$

Forecast differences outside the interval may be considered statistically significant for this chosen α. As shown, both the BIA and ETS of the Meso Eta are significantly different. But to what extent is the higher ETS of the Meso Eta a spurious effect of the difference in biases? As a proxy for correcting the model physics of one or the other forecast, we achieve similar biases here by adjusting the forecast precipitation thresholds of the NGM so that its BIA is similar to the Meso Eta BIA. For example, the NGM and the Meso Eta have similar biases at the 0.10-in. threshold if 0.09 in. replaces the 0.10-in. threshold for the NGM forecasts. In this manner, the NGM forecast threshold was adjusted at each observation threshold. The ETS was recalculated, and the hypothesis test methodology was reapplied. Results are shown in Fig. 3. As shown, the difference in ETS between the two forecast models is smaller but still significant after accounting for bias differences; the Meso Eta appears to be unambiguously better at precipitation forecasting than the NGM.

The resampling test demonstrated above is more computationally expensive and will take some time to code. Is a paired t test or Wilcoxon signed-rank test a reasonable substitute? Table 5 presents the p values of the

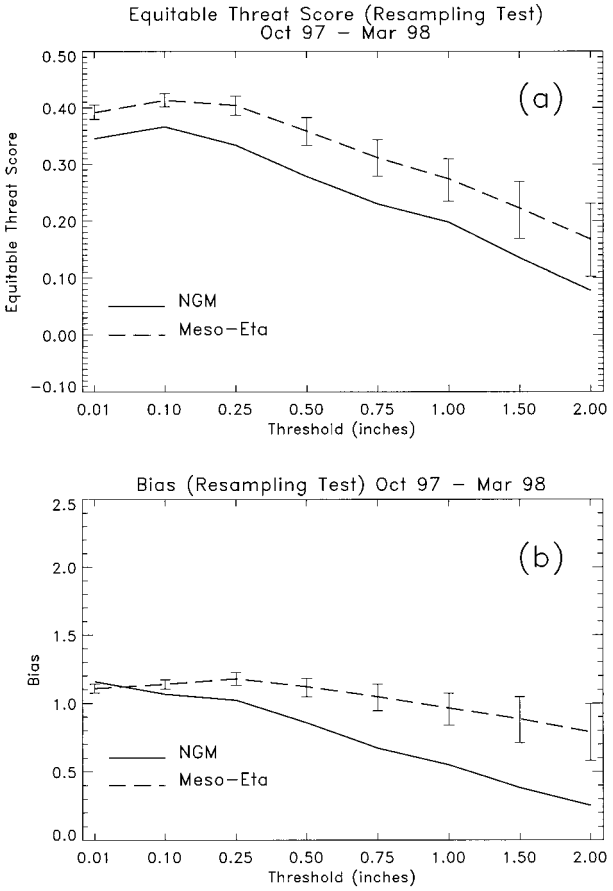

FIG. 2. (a) ETS and (b) BIA for a set of 160 forecasts of 24-h accumulated precipitation for Meso Eta and NGM forecasts. Overplotted error bars used to indicate 2.5th and 97.5th percentiles of resampled distribution, referenced to the Meso Eta forecast.

four tests at the various precipitation thresholds applied to the differences between the Meso Eta and bias-corrected NGM over the $n = 160$ case days. At lower precipitation thresholds all tests agree that differences in ETS are significant while differences in BIA are not. At higher thresholds, however, there are some notable differences among the hypothesis tests. For example, at 2.0 in., the resampling test for ETS yields a p value of $\sim$0.1, while the Wilcoxon and paired t tests still have p values near zero. For BIA, the Wilcoxon signed-rank test indicates a statistically significant difference in BIA at the 1.50-in. threshold even though the NGM's BIA was explicitly constructed to be the same as the Meso Eta's BIA (Fig. 3b). What may be affecting the test results at the higher thresholds? As previously discussed, only the resampling method in Eqs. (10)–(15) is designed to gracefully handle dry days, when there may be strong sensitivity to small changes in the contingency table population or ties among daily threat scores. In this resampling technique, the permutation of forecasts on the dry days will have little influence com-

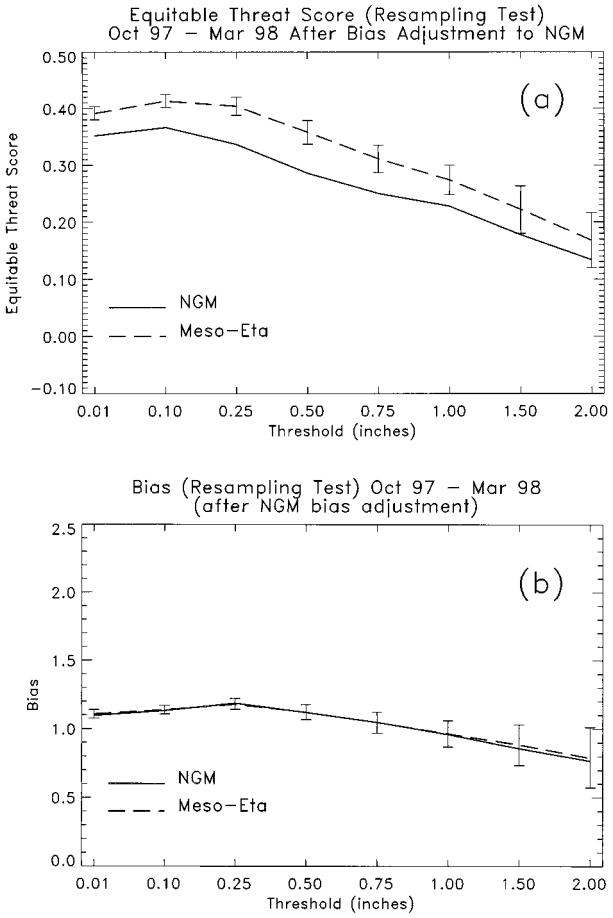

FIG. 3. As in Figs. 2a,b but after the NGM forecast thresholds were corrected to ensure a bias consistent with the Meso Eta bias.

pared to the permutation on the wet days. This is an important property; as shown in Fig. 4, there are many days where no events over the threshold were forecast to occur or verified, resulting in a tie in the daily threat score.

To determine more formally whether the apparent statistical significance of the *t* test and Wilcoxon test of ETS might be trusted, curves of the power of the tests were generated. The power measures the ''type II'' error of a test, that is, an incorrect acceptance of the null hypothesis when the alternative hypothesis is correct. Power is defined as 1 minus the probability of a type II error. Ideally, the power curve will be at α, the probability of a type I error (incorrect acceptance of alternative hypothesis) when the null hypothesis is true and grow quickly to 1.0 as the sample populations begin to differ. Here, power curves are generated by repeatedly conducting hypothesis tests using randomly generated contingency tables. For each case day, a contingency table (a, b, c, d) was randomly selected from one of the 320 Meso Eta or NGM forecasts. Next, a hypothetical perfect contingency table (a^*, b^*, c^*, d^*) was created on each day by assuming that all the forecast false alarms were hits, and all the misses were nonevents, that is, $a^* = a + b$, $b^* = 0$, $c^* = 0$, $d^* = c + d$, and $a + b + c + d = 1$. Given a desired fractional improvement x, where $x = 0$ is no improvement and $x = 1$ is a perfect forecast, the expected value of an improved contingency table $(\overline{a}_r, \overline{b}_r, \overline{c}_r, \overline{d}_r)_x$ is set, where

$$(\overline{a}_r, \overline{b}_r, \overline{c}_r, \overline{d}_r)_x = x(a, b, c, d) + (1 - x)(a^*, b^*, c^*, d^*). \quad (27)$$

Actual contingency tables were then randomly populated with 10 000 counts to have expectation (a, b, c, d)*10 000 and $(\overline{a}_r, \overline{b}_r, \overline{c}_r, \overline{d}_r)_x^*$10 000. The hypothesis

TABLE 5. The p values of hypothesis tests for differences in ETS and BIA between Meso Eta and bias-corrected NGM forecasts at various precipitation thresholds.

Threat score	Threshold (in.)	Resampling p value	Paired *t* test p value	Wilcoxon S-R p value	Resampling on daily threat score p value
ETS	0.01	<0.00001	<0.00001	<0.00001	<0.00001
	0.10	<0.00001	<0.00001	<0.00001	<0.00001
	0.25	<0.00001	<0.00001	<0.00001	<0.00001
	0.50	<0.00001	<0.00001	<0.00001	<0.00001
	0.75	<0.00001	<0.00001	<0.00001	<0.00001
	1.00	<0.00001	<0.00001	0.00065	<0.00001
	1.50	0.02002	<0.00001	0.00001	<0.00001
	2.00	0.10410	0.00018	<0.00001	<0.00001
BIA	0.01	0.240	0.292	0.124	0.494
	0.10	0.331	0.387	0.133	0.396
	0.25	0.360	0.399	0.181	0.045
	0.50	0.490	0.494	0.387	0.117
	0.75	0.492	0.498	0.484	0.092
	1.00	0.452	0.485	0.254	0.496
	1.50	0.365	0.415	0.004	0.272
	2.00	0.423	0.407	0.029	0.197

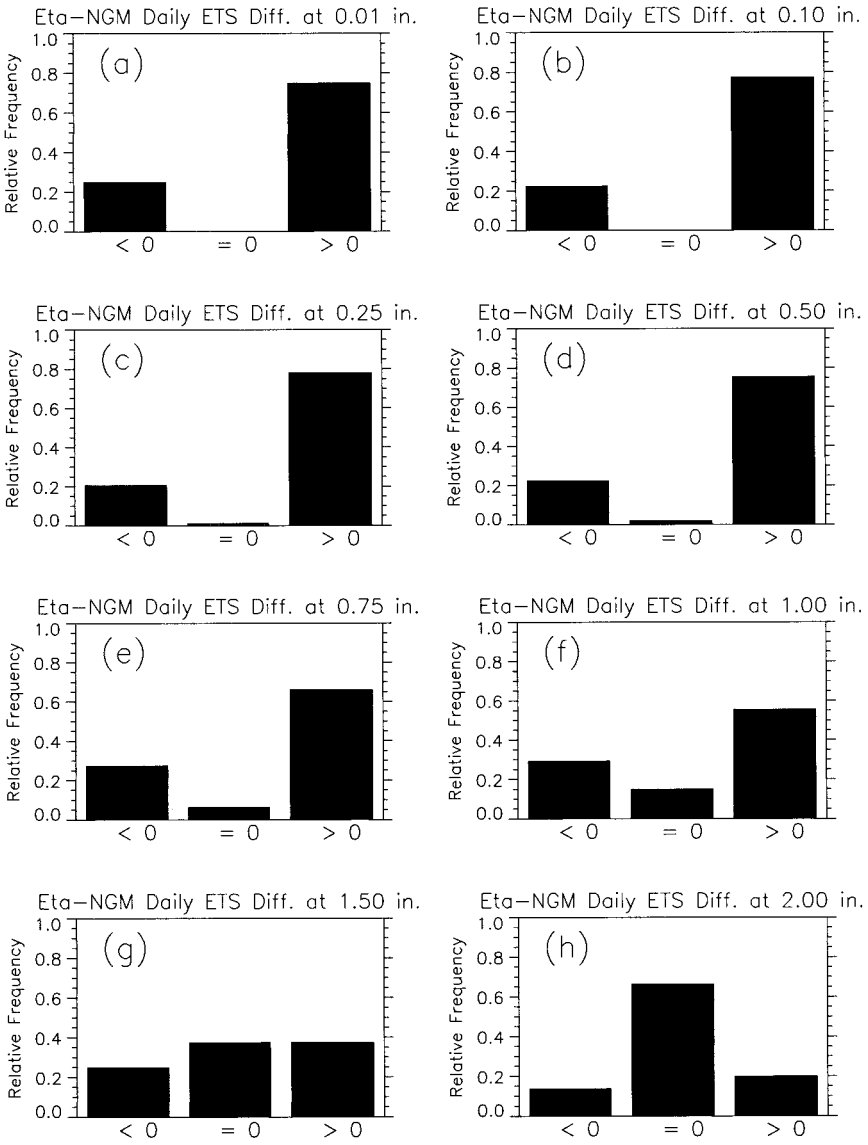

FIG. 4. Fraction of the case days where the domain-average Meso Eta ETS was greater than, equal to, or less than the NGM ETS: (a) 0.01-, (b) 0.10-, (c) 0.25-, (d) 0.50-, (e) 0.75-, (f) 1.00-, (g) 1.50-, and (h) 2.00-in. threshold.

test methodologies were applied, the acceptances/rejections of the null hypothesis tallied, and the process repeated 1000 times. From this, curves of the power were generated. Figures 5–7 show power curves for ETS evaluated at the 0.01-, 1.0-, and 2.0-in. thresholds. As shown, at 0.01 in. all tests perform similarly. As the precipitation threshold is increased and the event becomes rarer, the Wilcoxon test appears alternately more (1.0 in.) or less powerful (2.0 in.) than the other tests, especially for small sample sizes. The resampling test operating on the daily ETS is typically a powerful test, while the resampling test operating on contingency table elements is the least powerful. This illustrates that the apparent lack of power in this resampling test is an appropriate consequence of the designed insensitivity to small fluctuations in daily contingency table populations; resampling on daily threat scores produces results similar to the other two tests. Further, this hypothesis test methodology is the only one where the sample statistic is consistent with the way in which threat scores are calculated. Hence, it is the hypothesis test methodology of choice for threat scores.

b. Comparisons for probabilistic quantitative precipitation forecasts

Consider first how the resampling hypothesis test in (18)–(26) performed on the actual PQPFs. No statistical

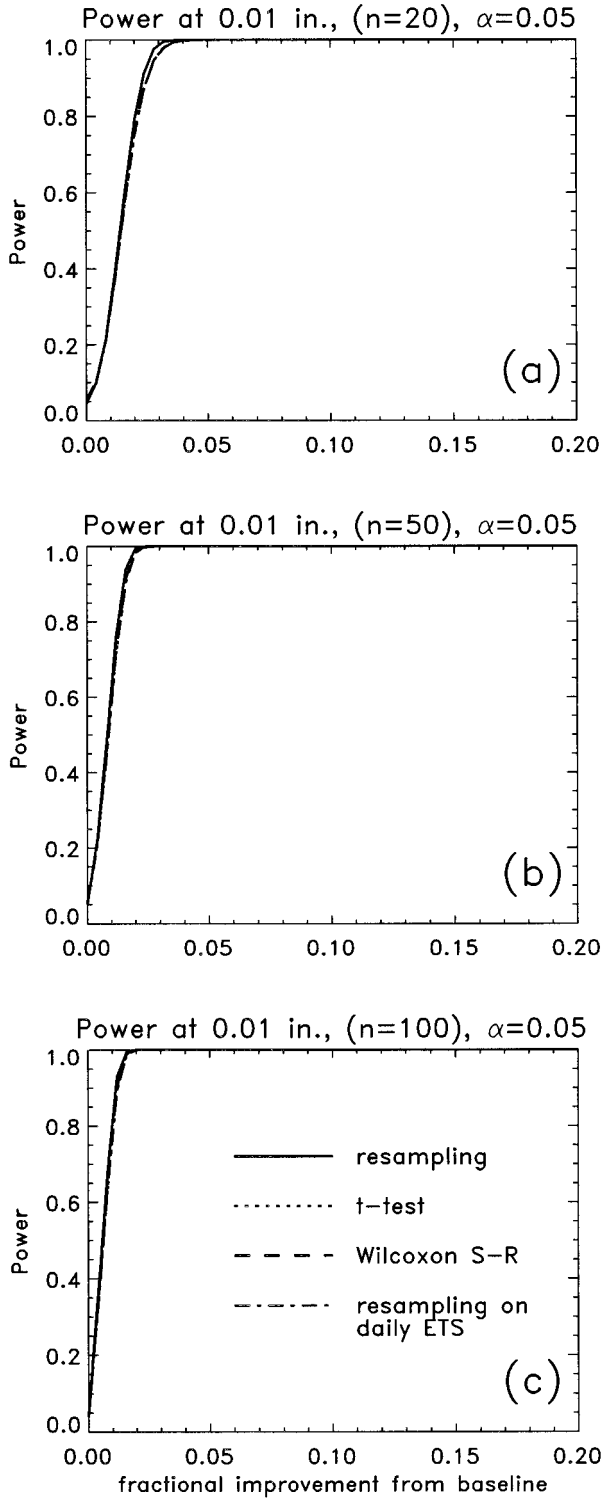

FIG. 5. Power curves for the resampling test, the paired *t* test, and the Wilcoxon signed-rank test on ETS for 0.01-in. threshold and a two-sided test with $\alpha = 0.05$: (a) $n = 20$, (b) $n = 50$, (c) $n = 100$.

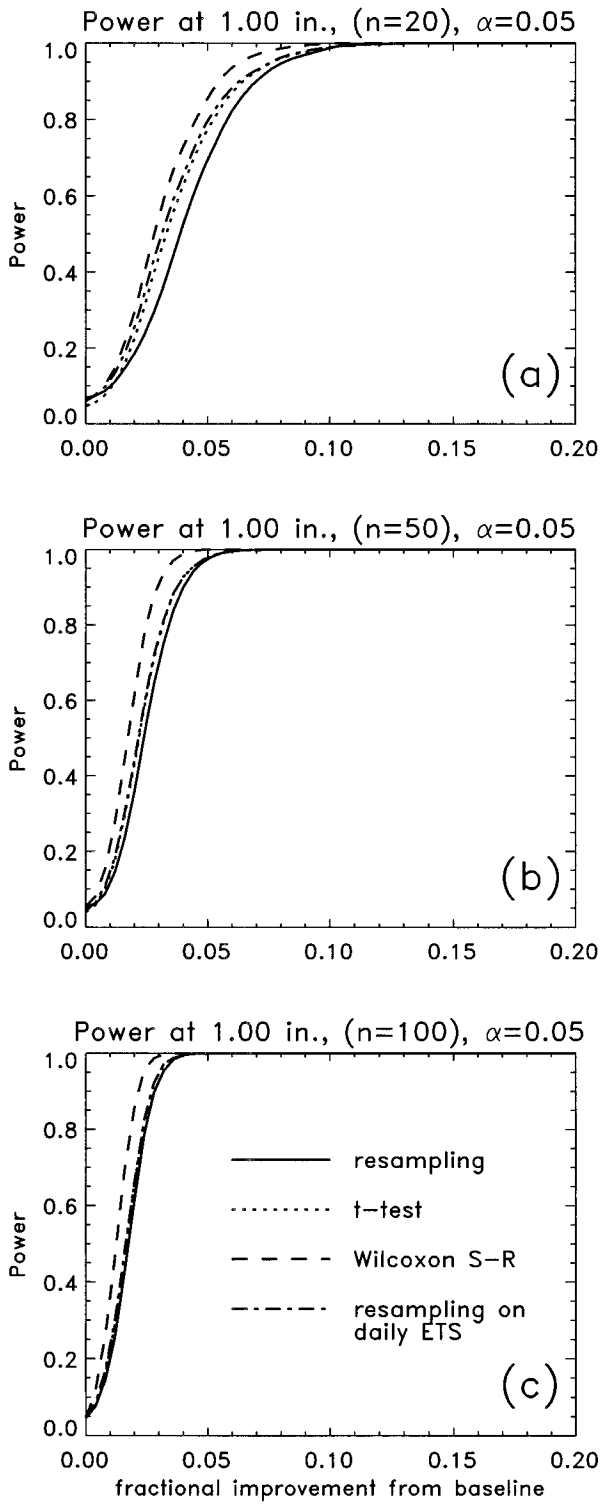

FIG. 6. As in Fig. 5 but for threshold = 1.0 in.

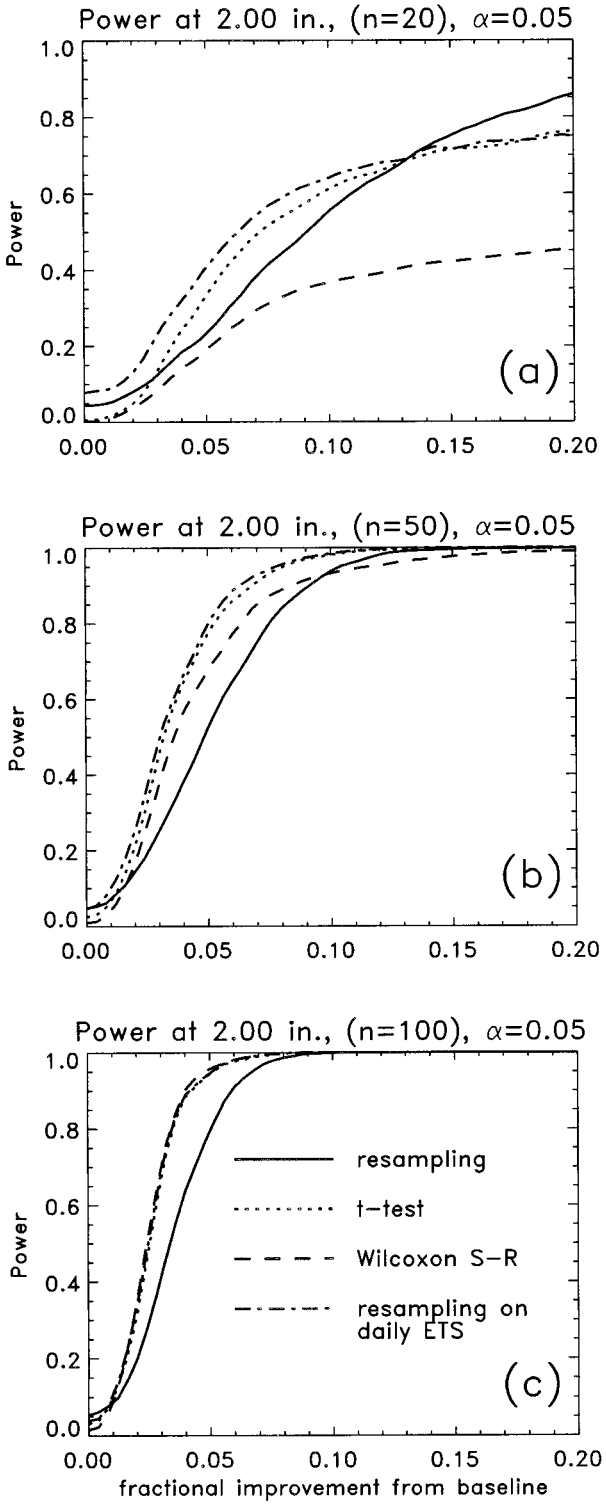

FIG. 7. As in Fig. 5 but for threshold = 2.0 in.

TABLE 6. The p values for hypothesis tests on difference in RPS between MOS and Γ ($\overline{\text{Eta}}$) forecasts.

Lead time (h)	Resampling	t test	Wilcoxon signed-rank
24	0.343	0.325	0.219
36	0.306	0.299	0.139
48	0.375	0.344	0.086

significance was noted for differences in RPSS between MOS and Γ($\overline{\text{Eta}}$) (Table 6). As shown, the location of the actual test statistic in the resampled distribution indicated no statistical significance for $\alpha = 0.05$. This indicates that either the model forecasts are truly indistinguishable with respect to RPSS or that at least testing must be performed on a larger sample. Here, the Wilcoxon or paired *t* tests are acceptable surrogates for the more complex resampling test (power curves not shown), but to be consistent with the way the RPSS is calculated, tests must be performed using the difference in daily sums of RPS rather than differences in daily RPSSs (Table 4).

5. Conclusions

The use of hypothesis tests to evaluate the significance of differences between competing precipitation forecasts is explored here. These tests should supplement but not replace the commonsense evaluation of forecast quality. Careful subjective evaluation and robust testing over a range of weather conditions will always be prudent, regardless of the significance of hypothesis tests.

Nonetheless, competing precipitation forecasts can be tested to evaluate whether improvements are statistically significant. A simple paired *t* test or Wilcoxon signed-rank test provides an estimate, but when evaluating threat scores these tests may be unduly sensitive to small changes in contingency table elements on dry days and are thus not recommended. The resampling technique operating on a vector of daily contingency table elements is preferred, since the methodology is insensitive to small changes in the contingency table population and is consistent with the way threat score statistics are calculated. For testing differences in RPSS, a simple Wilcoxon signed-rank test or *t* test is a worthy substitute for a resampling test, but the user should design the test to operate on the daily differences in sums of RPS rather than the daily differences in RPSS.

It is hoped that hypothesis testing will become more commonplace when evaluating the significance of model changes.

Acknowledgments. Barbara Hansford and Rick Katz are thanked for their review of this manuscript, and Mike Baldwin is thanked for supplying the forecast and verification data for the testing of threat scores. Bob Vislocky, Doug Nychka, and members of the Geo-

physical Statistics Program at NCAR are thanked for fruitful discussions on forecast verification problems. This research was done under an Advanced Studies Program postdoctoral fellowship at the National Center for Atmospheric Research, which is supported by the National Science Foundation.

REFERENCES

Black, T., 1994: The new NMC mesoscale eta model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

Briggs, W. M., and R. Levine, 1998: Comparison of forecasts using the bootstrap. Preprints, *14th Conf. on Probability and Statistics in the Atmospheric Sciences,* Phoenix, AZ, Amer. Meteor. Soc., 1–3.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting,* **4,** 401–412.

Daan, H., 1985: Sensitivity of the verification scores to classification of the predictand. *Mon. Wea. Rev.,* **113,** 1384–1392.

Diaconis, P., and B. Efron, 1983: Computer-intensive methods in statistics. *Sci. Amer.,* **248,** 116–130.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman and Hall, 436 pp.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Fritsch, J. M., and Coauthors, 1998: Meeting summary: Quantitative precipitation forecasts: Report of the Eighth Prospectus Development Team. U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.,* **79,** 285–299.

Good, P., 1994: *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer-Verlag, 228 pp.

Hall, P., and S. R. Wilson, 1991: Two guidelines for bootstrap hypothesis testing. *Biometrics,* **47,** 757–762.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta/RSM probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Hoke, J. E., and Coauthors, 1989: The regional analysis and forecast system of the National Meteorological Center. *Wea. Forecasting,* **4,** 323–334.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasting with the Eta regional model at the National Centers for Environmental Prediction. *Bull. Amer. Meteor. Soc.,* **77,** 2637–2649.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

Petersen, R. A., and Coauthors, 1991: Changes to NMC's regional analysis and forecast system. *Wea. Forecasting,* **6,** 133–141.

Rogers, E., D. G. Deaven, and G. J. DiMego, 1995: The regional analysis system for the operational ''early'' Eta Model: Original 80-km configuration and recent changes. *Wea. Forecasting,* **10,** 810–825.

——, and Coauthors, 1996: Changes to the operational ''early'' Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting,* **11,** 391–413.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Sprent, P., 1989: *Applied Nonparametric Statistical Methods.* Chapman and Hall, 519 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate,* **10,** 66–82.